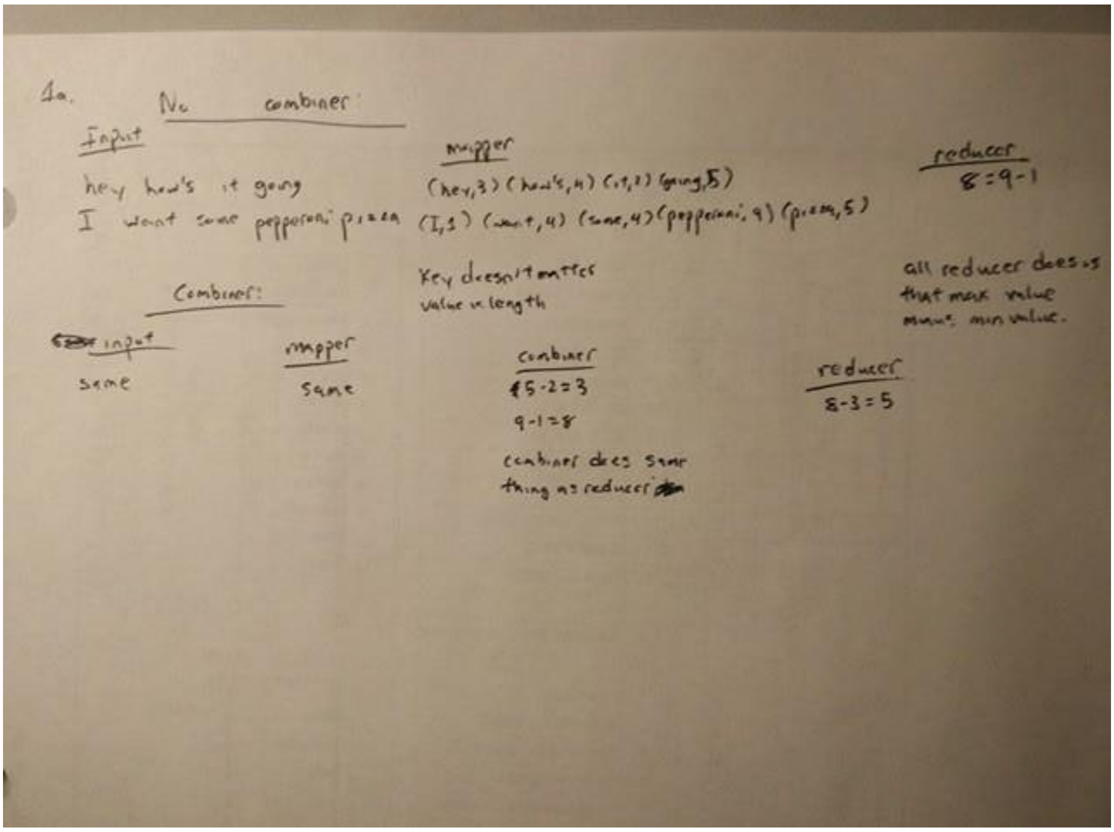
Terry Lyu 435091 wustl key: terrylu

Michael Wang 438275 wustl key: michaelwang

1a. You can use SumReducer as a Combiner for the WordCount problem because addition is both communicative and associative. An example of an instance where you cannot use a Reducer as a Combiner is if you were to try to find the difference between the longest word and the shortest word in a text file. Using only one reducer, you would simply sort the words by their length and subtract the length of the longest one by the length of the shortest one, and output that value, but if you were to use that reducer as a combiner, you would get the difference in length between the longest and shortest word per mapper node as output of the combiner. Then the reducer would output the difference between the largest difference and smallest difference in word length among the mapper nodes.

A example:



1a. No combiner

Input

hey how's it going
I want some pepperoni pizza

mapper

(hey,3) (how's,4) (it,2) (going,5)
(I,1) (want,4) (some,4) (pepperoni,9) (pizza,5)

Key doesn't matter
value is length

reducer

8=9-1

all reducer does is that max value minus min value.

Combiner:

input

same

mapper

same

combiner

5-2=3
9-1=8

combiner does same thing as reducer

reducer

8-3=5

b.
the command is job.setCombinerClass(SumReducer.class);

the result is the same as the original computation

```
To 1
be 2
is 1
not 2
or 1
that 1
to 1
```

partc

number of bytes read
residue:	10713042-769714=9943328

number of bytes written
residue: 21980312-2095251 = 19885061

map output records

residue : 964453-964453 = 0

combine input records

0-	964453 = -964453

Detail output please see the appendix

Part d

CPU time spent is 14580 without combiner, 18530 with combiner, physical memory snapshot is 1337413632 without combiner, 1285042176 with combiner.

It is a good idea to use combiner, because despite longer time is used by CPU, it still saved massive amount of memory. Which can be a key deciding factor for mapreduce jobs.

The time spent is higher when there is a combiner, because combiner needs to pre-process the data from mapper and then send it to the reducer, which will increase the time spent.
When the data we are trying to process(we need to transfer massive amount of data from mapper phase to reducer phase, and the function needs to be both associative and commutative) then implementing a combiner will not mess up the result, and save a great amount of memory.

Q2
C)

Positive line is 405, negative line is 805, the rest is 5215
The sensitivity score is -0.330579
The positive score is 0.334711

So Shakespeare’s poems are negative

d. these sentiment statistics are not the best way because we are using predefined word-based dictionaries to determine writer’s emotion. The dictionary could be not sufficient. Also it is not enough to determine the writer’s emotion entirely by words occurred in the sentence. The words might not be illustrating the writer’s personal emotions. So it could be helpful if we could expand the search using larger dictionaries, and even expand the search to phrase level or even sentence level.

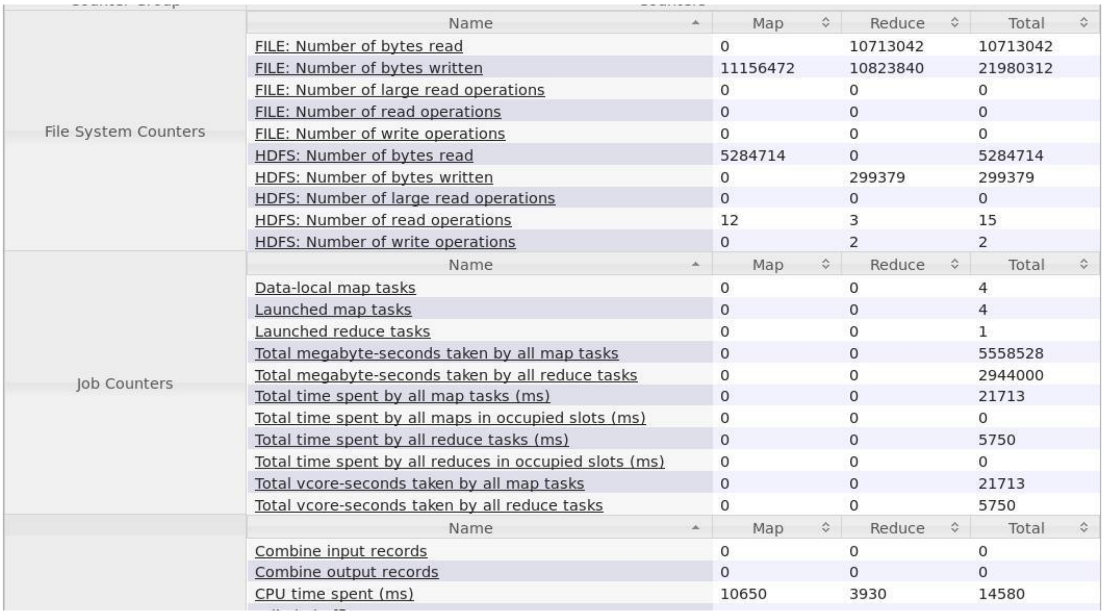
Q3

```
(Smith,John)     3
(Turing,Alan)     1
(Wamsley,Jayme)     1
(Webre,Josh)     1
(Weston,Clark)     1
(Woodburn,Louis)     1
(Woodburn,Providencia)     1
```

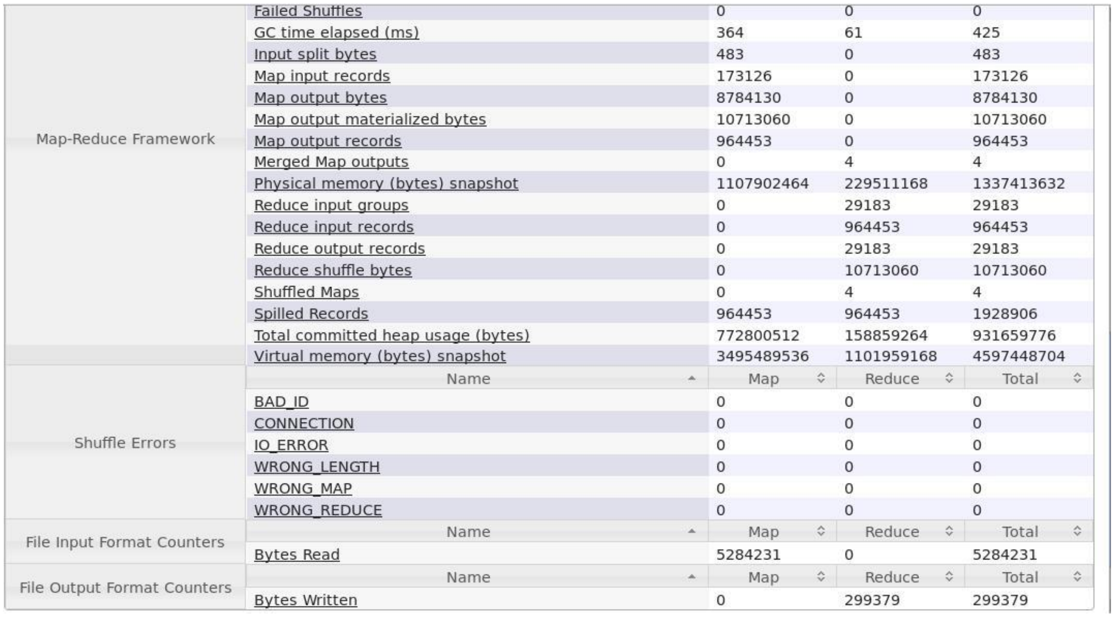
Terry Lyu 435091 wustl key: terrylu

Michael Wang 438275 wustl key: michaelwang

Q1 without combiner output



Counter Group	Name	Map	Reduce	Total
File System Counters	FILE: Number of bytes read	0	10713042	10713042
	FILE: Number of bytes written	11156472	10823840	21980312
	FILE: Number of large read operations	0	0	0
	FILE: Number of read operations	0	0	0
	FILE: Number of write operations	0	0	0
	HDFS: Number of bytes read	5284714	0	5284714
	HDFS: Number of bytes written	0	299379	299379
	HDFS: Number of large read operations	0	0	0
	HDFS: Number of read operations	12	3	15
	HDFS: Number of write operations	0	2	2
Job Counters	Name	Map	Reduce	Total
	Data-local map tasks	0	0	4
	Launched map tasks	0	0	4
	Launched reduce tasks	0	0	1
	Total megabyte-seconds taken by all map tasks	0	0	5558528
	Total megabyte-seconds taken by all reduce tasks	0	0	2944000
	Total time spent by all map tasks (ms)	0	0	21713
	Total time spent by all maps in occupied slots (ms)	0	0	0
	Total time spent by all reduce tasks (ms)	0	0	5750
	Total time spent by all reduces in occupied slots (ms)	0	0	0
	Total vcore-seconds taken by all map tasks	0	0	21713
	Total vcore-seconds taken by all reduce tasks	0	0	5750
	Name	Map	Reduce	Total
	Combine input records	0	0	0
	Combine output records	0	0	0
	CPU time spent (ms)	10650	3930	14580

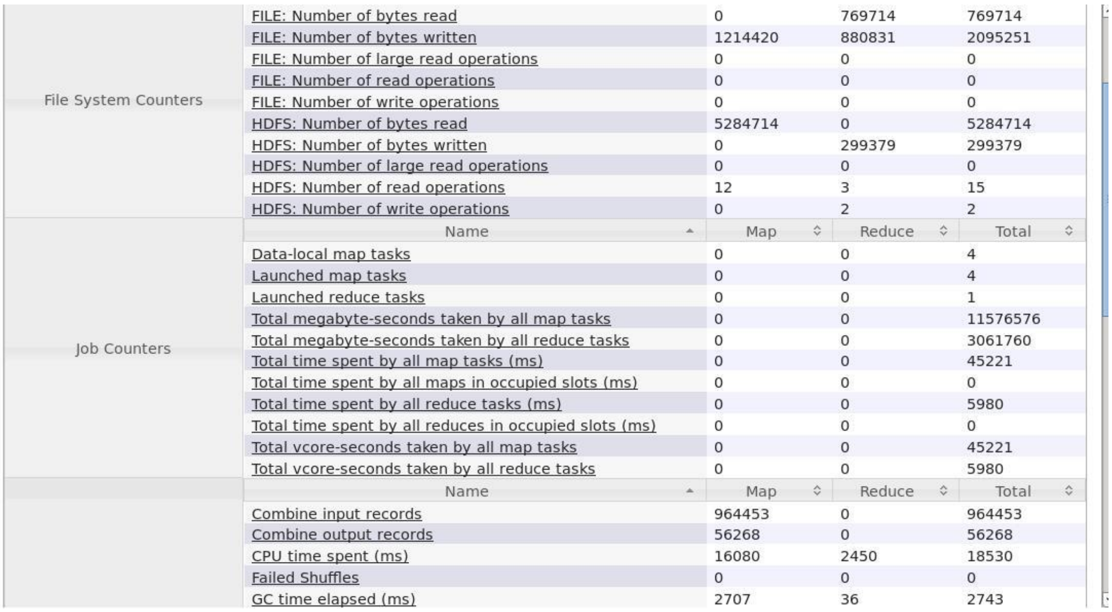


Counter Group	Name	Map	Reduce	Total
Map-Reduce Framework	Failed Shuffles	0	0	0
	GC time elapsed (ms)	364	61	425
	Input split bytes	483	0	483
	Map input records	173126	0	173126
	Map output bytes	8784130	0	8784130
	Map output materialized bytes	10713060	0	10713060
	Map output records	964453	0	964453
	Merged Map outputs	0	4	4
	Physical memory (bytes) snapshot	1107902464	229511168	1337413632
	Reduce input groups	0	29183	29183
	Reduce input records	0	964453	964453
	Reduce output records	0	29183	29183
	Reduce shuffle bytes	0	10713060	10713060
	Shuffled Maps	0	4	4
	Spilled Records	964453	964453	1928906
	Total committed heap usage (bytes)	772800512	158859264	931659776
	Virtual memory (bytes) snapshot	3495489536	1101959168	4597448704
Shuffle Errors	Name	Map	Reduce	Total
	BAD_ID	0	0	0
	CONNECTION	0	0	0
	IO_ERROR	0	0	0
	WRONG_LENGTH	0	0	0
	WRONG_MAP	0	0	0
	WRONG_REDUCE	0	0	0
File Input Format Counters	Name	Map	Reduce	Total
	Bytes Read	5284231	0	5284231
File Output Format Counters	Name	Map	Reduce	Total
	Bytes Written	0	299379	299379

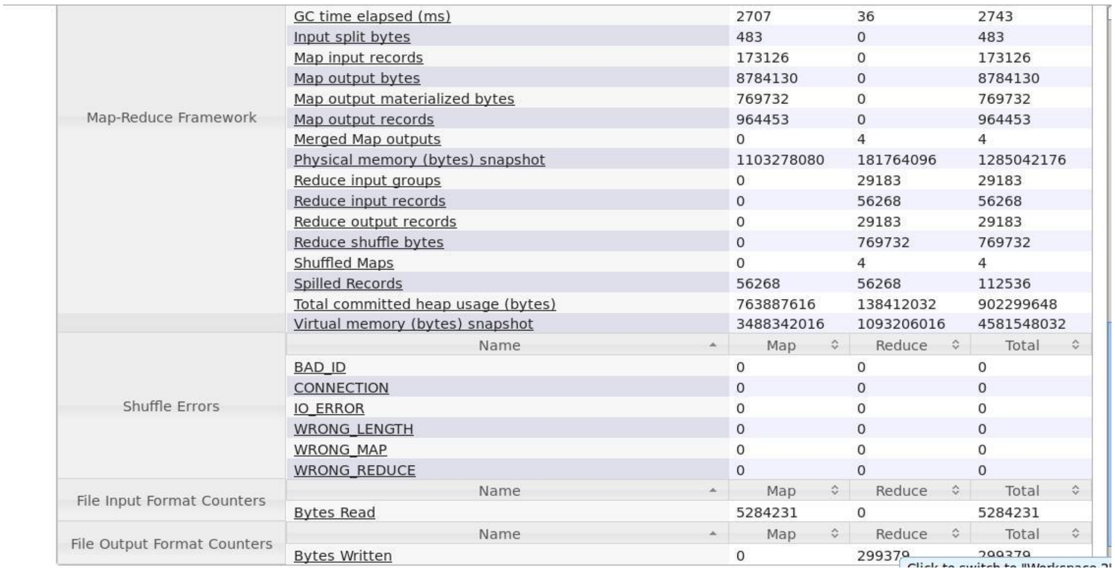
Terry Lyu 435091 wustl key: terrylu

Michael Wang 438275 wustl key: michaelwang

Q1 with combiner output



File System Counters	FILE: Number of bytes read	0	769714	769714
	FILE: Number of bytes written	1214420	880831	2095251
	FILE: Number of large read operations	0	0	0
	FILE: Number of read operations	0	0	0
	FILE: Number of write operations	0	0	0
	HDFS: Number of bytes read	5284714	0	5284714
	HDFS: Number of bytes written	0	299379	299379
	HDFS: Number of large read operations	0	0	0
	HDFS: Number of read operations	12	3	15
	HDFS: Number of write operations	0	2	2
Job Counters	Name	Map	Reduce	Total
	Data-local map tasks	0	0	4
	Launched map tasks	0	0	4
	Launched reduce tasks	0	0	1
	Total megabyte-seconds taken by all map tasks	0	0	11576576
	Total megabyte-seconds taken by all reduce tasks	0	0	3061760
	Total time spent by all map tasks (ms)	0	0	45221
	Total time spent by all maps in occupied slots (ms)	0	0	0
	Total time spent by all reduce tasks (ms)	0	0	5980
	Total time spent by all reduces in occupied slots (ms)	0	0	0
	Total vcore-seconds taken by all map tasks	0	0	45221
	Total vcore-seconds taken by all reduce tasks	0	0	5980
	Name	Map	Reduce	Total
	Combine input records	964453	0	964453
	Combine output records	56268	0	56268
	CPU time spent (ms)	16080	2450	18530
	Failed Shuffles	0	0	0
	GC time elapsed (ms)	2707	36	2743



Map-Reduce Framework	GC time elapsed (ms)	2707	36	2743
	Input split bytes	483	0	483
	Map input records	173126	0	173126
	Map output bytes	8784130	0	8784130
	Map output materialized bytes	769732	0	769732
	Map output records	964453	0	964453
	Merged Map outputs	0	4	4
	Physical memory (bytes) snapshot	1103278080	181764096	1285042176
	Reduce input groups	0	29183	29183
	Reduce input records	0	56268	56268
	Reduce output records	0	29183	29183
	Reduce shuffle bytes	0	769732	769732
	Shuffled Maps	0	4	4
	Spilled Records	56268	56268	112536
	Total committed heap usage (bytes)	763887616	138412032	902299648
	Virtual memory (bytes) snapshot	3488342016	1093206016	4581548032
Shuffle Errors	Name	Map	Reduce	Total
	BAD_ID	0	0	0
	CONNECTION	0	0	0
	IO_ERROR	0	0	0
	WRONG_LENGTH	0	0	0
	WRONG_MAP	0	0	0
	WRONG_REDUCE	0	0	0
File Input Format Counters	Name	Map	Reduce	Total
	Bytes Read	5284231	0	5284231
File Output Format Counters	Name	Map	Reduce	Total
	Bytes Written	0	299379	299379